BMJ Open

Increased risk of COVID-19 related admissions in cancer patients in the West Midlands region of the United Kingdom: A Retrospective cohort study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-053352
Article Type:	Original research
Date Submitted by the Author:	11-May-2021
Complete List of Authors:	Akingboye, Akinfemi; Russells Hall Hospital Mahmood, Fahad; Russells Hall Hospital Amiruddin, Nabeel; Russells Hall Hospital Reay, Michael; Russells Hall Hospital Nightingale, Peter; University Hospitals Birmingham NHS Foundation Trust, Institute of Translational Medicine Ogunwobi, Olorunseun; Hunter College CUNY, Biological Sciences
Keywords:	COVID-19, Risk management < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, ONCOLOGY

Title page

Title: Increased risk of COVID-19 related admissions in cancer patients in the West Midlands region of the United Kingdom: A Retrospective cohort study

Authors:

Akinfemi Akingboye (MD, FRCS)[1], Fahad Mahmood (PhD, MRCS)[1], Nabeel Amiruddin (FRCA)[1], Michael Reay (FRCA)[1], Peter Nightingale[2], Olorunseun Ogunwobi (MBBS, MSc, PhD)[3,4]

1. Russells Hall Hospital, Dudley, United Kingdom, DY1 2HQ
2. Institute of Translational Medicine, WCL - University Hospitals Birmingham NHS Foundation Trust
3. Hunter College Center for Cancer Health Disparities Research (CCHDR), New York
4. Hunter College of The City University of New York

Corresponding authors:

Olorunseun Ogunwobi, MBBS, MSc, PhD

Director, Hunter College Center for Cancer Health Disparities Research (CCHDR)

Associate Professor of Biological Sciences

Hunter College of The City University of New York

Belfer Research Building, Room 426, 413 E 69th Street, New York, NY 10021

Tel: 212-896-0447

E-mails: oo158@hunter.cuny.edu

Mr Akinfemi Akingboye MBBS, MD (Lond), FRCS (Gen. Surg.)

Consultant Laparoscopic Colorectal & General Surgeon

Department of General Surgery

The Dudley Group NHS Trust,

Russells Hall Hospital Dudley, DY1 2HQ

Email: a.akingboye@nhs.net

Tele: 01384 456111 Ext 2739

Word count: 2829

Abstract

Objective: Susceptibility of cancer patients to COVID-19 pneumonitis has been variable. We aim to quantify the risk of hospitalization in active cancer patients and use a machine learning algorithm (MLA) and traditional statistics to predict clinical outcomes and mortality.

Design: Retrospective cohort study.

Setting: A single United Kingdom(UK) district general hospital (Rec-Reference 20/EE/0139;IRAS ID28233).

Participants: Data on total hospital admissions between March 2018-June 2020, all active cancer diagnoses between March 2019-June 2020 and clinical parameters of COVID-19 positive admissions between March 2020 and June 2020 were collected. 526 COVID-19 admissions without an active cancer diagnosis were compared with 87 COVID-19 admissions with an active cancer diagnosis.

Primary & secondary outcome measures: 30 and 90-day post-COVID-19 survival.

Results: In total 613 patients were enrolled with Male to Female ratio of 1:6 and median age 77 years. The estimated infection rate of COVID-19 was 87/22729 (0.4%) in the cancer patients and 526/426658 (0.1%) in the non-cancer population (OR:3.105; 95%CI 2.474-3.897; p<0.001). Survival was reduced in cancer patients with COVID-19 at 90 days. R-Studio software determined the association between cancer status, COVID-19 and 90-day survival against variables using MLA. Multivariate analysis showed increases in age (OR1.039[95%CI1.020-1.057], p<0.001), urea (OR1.005[95%CI1.002-1.007],p<0.001) and CRP (OR1.065[95%CI1.016-1.116],p<0.008) is associated with greater 30-and 90-day mortality. The

MLA model examined the contribution of predictive variables for 90-day survival (AUC: 0.749); with transplant patients, age, male gender and diabetes mellitus being predictors of greater mortality.

Conclusions: Active cancer diagnosis has a 3-fold increase in risk of hospitalization with COVID-19. Increased age, urea, and CRP predict mortality in cancer patients. MLA complements traditional statistical analysis in identifying prognostic variables for outcomes of COVID-19 infection in cancer patients. This study should inform a redesign of cancer services to ensure safe delivery of cancer care.

Trial Registration: Not applicable. Observational study, not a clinical trial.

Strength & Limitations of this Study

- The study uses novel analytical methods derived from machine learning to evaluate risk from COVID-19 in cancer patients from hospitalization to mortality.
- Statistical and machine learning methods are compared to develop a profile of factors that can worsen outcomes from COVID-19 in cancer patients.
- The study analyses COVID-19 outcomes in cancer patients in a cohort covering a single UK metropolitan region only.
- Patients with COVD-19 and cancer who did not require admission to hospital were not included in this study.

Funding Statement: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. However, Olorunseun O. Ogunwobi is supported by National Cancer Institute grant # CA221704.

Competing interests statement: No conflicts of interest to declare.

Introduction

The severe acute respiratory syndrome coronavirus 2 leads to the coronavirus disease 2019 (COVID-19) (1,2). This highly transmissible disease has led to a global pandemic contributing to significant morbidity and mortality. Increased susceptibility and severity of COVID-19 are attributed to increasing age, smoking status, chronic obstructive pulmonary disease, diabetes mellitus, obesity, cardiovascular disease as well as cancer (3–6). In addition, the prevalence of all types of active or previous cancer in the United Kingdom (UK) is reported at 2.5 million cases with an incidence of 1000 newly diagnosed cases each day (7). Increased susceptibility to COVID-19 in cancer patients has been attributed to immune suppression and cancer treatments such as cytotoxic chemotherapy and immunotherapy (8,9). However, it is still not established whether this translates into increased hospitalization, illness severity or mortality risk. Risk adjusted models quote a mortality risk of between 25-39% in cancer patients hospitalized with COVID-19 (10). With increasing prevalence of COVID-19 in the UK, the impact of cancer on COVID-19 remains an area of active concern (8).

In addition, machine learning has become increasingly applied in healthcare settings to generate predictive models from a large number of variables (11–13). This distinguishes machine learning from traditional statistical modelling, such as regression analysis, by its ability to perform non-linear modelling utilizing large volume datasets and greater number of variables from registries (11,12). This can enable prognostic and diagnostic models to inform healthcare decision making.

This study aims to quantify the risk of hospitalization in active cancer patients as well as specific differences in clinicopathological and biochemical parameters between cancer and non-cancer COVID-19 patients using a machine learning algorithm. This will inform the recalibration of cancer services to ensure the highest quality of care for these patients during the pandemic.

Methods

A single UK centre retrospective cohort study was conducted **(Rec-Reference 20/EE/0139; IRAS ID28233).** Ethical approval was obtained from the Russells Hall ethics committee. Data on total hospital admissions between March 2018 and June 2020 was obtained from the local Information Technology (IT) department with a record of all hospital admission before and during the COVID-19 pandemic linking this with the Somerset cancer database to extrapolate the total number of active cancer patients that were admitted during the study period. Furthermore, all active solid organ cancer diagnoses between March 2019 and June 2020 were obtained from the local cancer network. This was used to determine the total number of active cancer patients with COVID-19 with the denominator being the total and active cancer population in the Dudley, West Midlands (UK) region. Biochemical and haematological parameters in the first 48 hours of admission along with 30 and 90-day post-COVID-19 survival was determined.

Patients below the age of 18 years and those with non-solid organ cancers were excluded. Moreover, patients who attended the emergency department and were not admitted were also excluded. COVID-19 diagnosis was established with a positive Reverse Transcriptase Polymerase Chain reaction (RT-PCR) test from an oropharyngeal swab Criteria for admission to hospital and critical care were determined by individual clinical assessment and oxygen requirement as well as ventilatory support. Data security was maintained through the RedCap uploading system.

Binary logistic regression analyses with survival status at 30 days as the dependent variable were used to estimate the univariable association with mortality for each explanatory variable. Age-adjusted associations were calculated in a similar way by including age at admission as a continuous variable in each model after checking the assumption of a linear effect of age on the log odds. Both forward and backward stepwise methods were used to determine the final multivariable model. These analyses were performed with SPSS 25.0.

Machine Learning Algorithm: Data pre-processing

R-Studio software was used to determine the association between cancer status, COVID-19 and 90-day survival against variables in a Machine Learning Algorithm. The conduct and reporting of our MLA was done in accordance with best practice guidance (11).

Feature selection:

The proportion of missing data was calculated for each variable and variables with less than 40% missing data were included in the analysis. This resulted in 33 variables being included for imputation of missing data, further pre-processing and model development.

All data within the Gender variable was replaced with ‘F’(female) and ‘M’(male). The documented ethnicities were replaced with the 3 categories of ‘European’, ‘South-Asian’ or ‘Afro-Caribbean’. The blood pressure information was split into systolic and diastolic pressures. A new dummy variable of ‘Mean arterial pressure’ was derived from the estimate: (Diastolic pressure + (Pulse pressure/3)). The pure numeric values from the entered data for oxygen saturations were extracted. For example, 97% would be changed to 97. A dummy variable was created from the difference in time between the date of onset of symptoms and date of hospital admission. This time-interval was recorded in days.

Data partition:

The overall dataset was partitioned into training and test sets. The training set was used purely for model training and hyperparameter tuning. The test set would only be used for model evaluation against new data. Partitioning was by a random allocation, while ensuring an identical distribution of patients who died at 90 days between both training and test sets. 75% of patients were allocated to the training set, with the remaining going into the test set.

Imputation of missing values:

Missing values were replaced with predicted values using k-nearest neighbour's model. This method designated the variable of a missing value as an outcome variable within a predictive model. A prediction of the missing value on the most similar k number of patients based on their other variables. The value k is a hyperparameter which was set to 10 after comparing the values of 5 and 10 without any difference. This imputation was performed separately on the training and test datasets in order to minimise overfitting of the final model by having the training dataset influence the imputation of values into the test dataset.

Additional pre-processing:

The degree of skewness of each the distribution of each variable was determined. A skewness value of 0 represented an absence of any skew, while a positive and negative value represents a positive and negative skew respectively. The following transformations were implemented before recalculating skewness:

1. Square-root
2. Log10
3. Reciprocal (1/x)

The transformations resulting in the least skews were implemented for each variable.

The data was then scaled such that the values of each variable had a mean of 0 and a standard deviation of 1.

Model development:

The caret package of R was called. Repeated 10-fold cross-validation was chosen (3 repetitions). Hyperparameter tuning of models was to optimised a metric of area under the receiver-operator curve.

The following models were trained: (1) Logistic regression, (2) regularised general linear model (3) k-nearest neighbours (4) Random forest (5) Neural network with 1 hidden. The resultant models were compared by their respective areas under the receiver-operator curves.

Results

In total, 22,729 active cancer patients were identified in the Dudley West Midlands region out of a catchment size of 426658 patients in the region from the local cancer network. 87/22729 (0.4%) cancer patients in the Dudley region were admitted with COVID-19 compared with 526/426658 (0.1%) COVID-19 admissions without cancer during the study period (Hazard ratio:3.105; 95%CI 2.474-3.897; p<0.001). Thus, the risk of hospital admission on presentation with COVID-19 increased 3-fold in the presence of an active cancer diagnosis.

Excluding those with incomplete data, the mean age of cancer patients was 77.8 (sd=12.3) years compared to 70 (sd=17.5) years (t-test; p<0.001). The Male:Female ratio was similar between the two groups. The majority of patients were of Caucasian ethnicity with similar distribution of diabetes, cardiovascular disease, transplant recipient and smoking status. Moreover, the median white cell count and CRP were similar between cancer and non-cancer patients. Thus, both cancer and non-cancer groups affected by COVID-19 had similar baseline characteristics. This is summarized in Table 1.

Table 1: Comparing characteristics of cancer and non-cancer patients. Values are counts and percentages except where stated. The p values are from Fisher's exact test, except for age (from a t test), white cell count and CRP (both from Mann-Whitney tests).

	CANCER PATIENTS (N=80)	NON-CANCER PATIENTS (N=276)	P VALUE
AGE IN YEARS (N=356) : MEAN (SD)	77.8 (12.3)	70.0 (17.5)	<0.001
SEX			0.699
FEMALE	34 (43%)	112 (41%)	
MALE	45 (57%)	164 (59%)	
ETHNICITY			0.280
AFRO-CARIBBEAN	1 (1%)	9 (4%)	
EUROPEAN	70 (95%)	197 (87%)	
SOUTH ASIAN	3 (4%)	50 (9%)	
SMOKING			0.176
CURRENT	5 (22%)	18 (30%)	

EX	9 (39%)	11 (18%)	
NEVER	9 (39%)	31 (52%)	
CARDIOVASCULAR			0.103
YES	22 (31%)	107 (41%)	
NO	50 (69%)	152 (59%)	
DIABETES MELLITUS			0.885
YES	22 (29%)	73 (28%)	
NO	53 (71%)	187 (72%)	
TRANSPLANT PATIENT			0.644
YES	2 (3%)	5 (2%)	
NO	70 (97%)	261 (98%)	
REASON FOR ADMISSION			<0.001
YES	48 (70%)	227 (90%)	
NO	21 (30%)	24 (10%)	
WHITE CELL COUNT (N=332) : MEDIAN (LOWER QUARTILE – UPPER QUARTILE)	8.8 (5.6-12.7)x10^9/L	7.2 (5.3-10.6)x10^9/L	0.096
CRP (N=324) : MEDIAN (LOWER QUARTILE – UPPER QUARTILE)	77 (22-135) mg/L	84 (36-157) mg/L	0.115

The machine learning algorithm examined the relative contributions of predictive variables on 90-day survival (Figure 1). We derived a generalized linear model with an area under the curve (AUC) of 0.749 to identify variables that predicted mortality from COVID-19 at 90 days in patients with active cancer (Figure 2) with transplant patient, age, gender and diabetes mellitus status being the most predictive in determining outcome from COVID-19. Since we accepted variables with up to 40% missing values (Supplementary Figure 1), imputation was performed using a separate k-nearest neighbours' algorithm, whereby a prediction of a missing value was made based the other available values, having been trained on the other patient data.

Our initial age-adjusted univariate analysis identified age, CRP, urea, creatinine, eGFR, haemoglobin and low initial blood pressure as significantly correlating with mortality risk (Supplementary Table 1). A further multivariate analysis of 33 out of 213 clinical variables with >60% data completeness showed increased age (Hazard ratio 0.915 [95%CI 0.870-0.960], p<0.001), urea (Hazard ratio 1.005 [95%CI

1.002-1.007], p<0.001) and CRP (Hazard ratio 1.065 [95%CI1.016-1.116], p<0.001) to be associated with greater risk of 30-and 90-day mortality (Table 2).

Table 2: Multivariate analysis showing increased age, CRP and urea are associated with the highest 90-day mortality risk in COVID-19 patients.

	P-valve	***OR***	***Lower CI***	***Upper CI***
Age	0.000	1.039	1.020	1.057
CRP	0.001	1.005	1.002	1.007
Urea	0.008	1.065	1.016	1.116

Kaplan-Meier survival analysis revealed reduced overall survival for patients with COVID-19 and cancer (Figure 3). However, Log-Rank analysis did not show significant difference between cancer and non-cancer COVID-19 patients (Log-rank p=0.172).

Discussion

Our study demonstrates that the presence of active cancer increased by 3-fold the risk of hospitalization with COVID-19. Moreover, higher CRP and urea are associated with greater mortality at 30- and 90-days post-diagnosis of COVID-19. These findings show that cancer patients who develop COVID-19 are likely to have a more severe form of the infection that would require supportive care in hospital. It also provides tools for monitoring patient response to treatment with high urea and CRP being poor prognostic markers and likely a consequence of severe COVID-19 This has implications for how we can deliver safe care to cancer patients in the ongoing pandemic as well as emerging from it given the restrictions on cancer services.

Several studies have reported prevalence and mortality risk of COVID-19 in cancer patients with a systematic review by Zarifkar *et al* identifying 110 studies covering 10 countries (14). The pooled prevalence of active cancer in COVID-19 positive hospitalized patients was 2.6% (95% CI 1.8-3.5%) across 37 cohort studies. Furthermore, there was a noticeable difference in the prevalence between western countries (5.6%, 95% CI 4.5%-6.7%) and China (1.7%, 95% CI 1.3%-2.3%) reflecting the underlying cancer prevalence. In addition, in-hospital mortality of 14.1% (95% CI 9.1-19.8%) for cancer and COVID-19 was derived from 17 retrospective cohort studies covering 904 patients (14). The mortality rate of 12.6% in a Brazilian cohort was also similarly reported (15). This indicated that COVID-19 patients with cancer had a 5-fold greater risk of death compared with non-cancer patients without other co-morbidities (14,16). However, there was significant heterogeneity between these studies (I^2=55.9%, P<0.01) with the type of cancer, stage and treatment regimen only specified in 8 studies along with incomplete followup. Furthermore, Liang et al reported a 28% prevalence of lung cancer amongst hospitalized cancer patients with COVID-19 (9). This reflects higher COVID-19 mortality rates in specific cancer patients including lung and haematological malignancy (9,14,15). Further studies have reported 3.5 fold increase in ICU admission or need for mechanical ventilation in COVID-19 patients with cancer (9). Using admission risk as a surrogate marker of severity in COVID-19, our results are consistent with the literature showing a 3-fold higher risk of admission with COVID-19 in the presence of cancer which will likely impact the delivery of care to these particular subgroups of patients.

There is likely to be a surge in demand for cancer services as well as predicted poor long-term survival in cancer patients due to delays in diagnosis and treatment (11,17). Over the first national UK lockdown, there was a 84% reduction in urgent cancer referrals which modelling predicted would lead to 181 additional lives lost or 3316 life-years lost with an average presentation delay of 2 months per patient (18). Although having cancer puts patient at increased risk of hospitalization with COVID-19, this must be balanced against risks of delayed treatment leading to disease progression to incurable stages (19). Particular cancers where timely intervention is critical such as pancreatic, lung and haematological malignancy should not have delays to treatment whereas others including prostate and non-melanoma skin cancers treatment may be safely delayed in selected patients (19). Several strategies including: Delays to surgery or chemotherapy, switching to oral or monotherapy treatment regimens, strict infection control protocols, online consultation, use of hypofractionated radiotherapy and provision of ITU support to these patients is essential to mitigate risk (10,19–21). This may be supplemented where possible with COVID-19 free 'cold' sites to reduce risk of transmission and prevent anti-cancer treatment induced COVID-19 (21). Thus categorization of patients according to risk, minimizing patient exposure and considering alternative regimens to control cancer forms the basis of current recommendations including the ESMO expert consensus and UK NHS guidelines (4,21). Furthermore, this data does not support delays in cancer treatment to reduce risk of COVID-19 transmission in cancer patients.

Several biochemical markers have been associated with a severe COVID-19 disease course. Zhou *et al* identified a raised D-dimer above 1μg/ml to be associated with a higher mortality risk (22). Furthermore, they identified low albumin, raised LDH, troponin, ferritin and IL-6 to be more prevalent in non-survivors. In addition, raised CRP and low glomerular filtration rate was associated with a more severe disease outcome with 18% deaths recorded in a renal transplant cohort in keeping with the severe disease course predicted in this group of immunocompromised patients (23).Our model identified raised urea and CRP in addition to transplant status as predictors of greater mortality risk which may lower threshold for admission or earlier referral for intensive care support. However, our algorithm could not specify the direction or size of this interaction which is a limitation of such models.

Machine learning algorithms are increasingly being used to support healthcare applications. It can be used to learn from data and identify hidden patterns. This can further help with decision-making processes in a variety of healthcare applications (24). Nonetheless, it requires training, validation and testing datasets to establish internal and external validity. Moreover, we have shown through our modelling that the findings of both MLA and traditional statistical analysis are complementary and may be used to generate a risk prediction scoring system in cancer patients with COVID-19. However, there are several limitations in the data presented. Confounders including smoking status or respiratory co-morbidity were not assessed which could influence outcomes in cancer patients. Patients with active cancer who tested positive for COVID-19 in the community but did not require hospital admission could not be evaluated. Having a broad inclusion criterion with all solid organ cancers whilst beneficial for looking at overall impact on cancer patients does not capture the granularity of how individual cancers may differ in their impact on COVID-19 patients. Our dataset was underpowered to perform relevant sub-group analyses on these patients. Although all active cancer patients were analysed, variation in the stage of cancer and treatment protocols were not accounted. Moreover, the machine learning algorithms are limited by missing data and rely upon imputation as part of model development. This needs external validation once developed which we have not performed.

Conclusions

COVID-19 has impacted both individuals and healthcare systems in an enormous way. How we deliver safe and effective care to these patients in the confines of our healthcare systems is predicated on identifying those most as risk from this disease. Machine learning algorithms provide an additional tool to for risk assessment to delineate factors with poor prognosis. This will enable us to reconfigure our healthcare systems to provide safe care to these more vulnerable patients.

References

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet (London, England). 2020 Feb;395(10223):497–506.

2. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. First Case of 2019 Novel Coronavirus in the United States. N Engl J Med. 2020 Mar;382(10):929–36.

3. Emami A, Javanmardi F, Pirbonyeh N, Akbari A. Prevalence of Underlying Diseases in Hospitalized Patients with COVID-19: a Systematic Review and Meta-Analysis. Arch Acad Emerg Med. 2020;8(1):e35.

4. Curigliano G, Banerjee S, Cervantes A, Garassino MC, Garrido P, Girard N, et al. Managing cancer patients during the COVID-19 pandemic: an ESMO multidisciplinary expert consensus. Vol. 31, Annals of oncology : official journal of the European Society for Medical Oncology. 2020. p. 1320–35.

5. Jordan RE, Adab P, Cheng KK. Covid-19: risk factors for severe disease and death. Vol. 368, BMJ (Clinical research ed.). England; 2020. p. m1198.

6. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA. 2020 May;323(20):2052–9.

7. MacMillan Trust. Macmillan Cancer Support Cancer in numbers. [Internet]. Macmillan Cancer Support. 2020. Available from: https://www.macmillan.org.uk/about-us/media-centre/facts-and-figures/cancer-in-numbers.html

8. UKCCM. The UK Coronavirus Cancer Monitoring Project: protecting patients with cancer in the era of COVID-19. Lancet Oncol. 2020 May;21(5):622–4.

9. Liang W, Guan W, Chen R, Wang W, Li J, Xu K, et al. Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China. Lancet Oncol. 2020 Mar;21(3):335–7.

10. Abdihamid O, Cai C, Kapesa L, Zeng S. The Landscape of COVID-19 in Cancer Patients: Prevalence, Impacts, and Recommendations. Cancer Manag Res. 2020;12:8923–33.

11. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016 Dec;18(12):e323.

12. Scott IA. Demystifying machine learning - a primer for physicians. Intern Med J. 2021 Jan;

13. Sarijaloo F, Park J, Zhong X, Wokhlu A. Predicting 90 day acute heart failure readmission and death using machine learning-supported decision analysis. Clin Cardiol. 2021 Feb;44(2):230–7.

14. Zarifkar P, Kamath A, Robinson C, Morgulchik N, Shah SFH, Cheng TKM, et al. Clinical Characteristics and Outcomes in Patients with COVID-19 and Cancer: a Systematic Review and Meta-analysis. Clin Oncol (R Coll Radiol). 2021 Mar;33(3):e180–91.

15. Fernandes GA, Feriani D, França e Silva ILA, Mendonça e Silva DR, Arantes PE, Canteras J da S, et al. Differences in mortality of cancer patients with COVID-19 in a Brazilian cancer center. Seminars in Oncology. 2021.

16. Ioannidis JPA, Axfors C, Contopoulos-Ioannidis DG. Population-level COVID-19 mortality risk for non-elderly individuals overall and for non-elderly individuals without underlying diseases in pandemic epicenters. Environ Res. 2020 Sep;188:109890.

17. Oncology TL. COVID-19: global consequences for oncology. Vol. 21, The Lancet. Oncology. 2020. p. 467.

18. Sud A, Torr B, Jones ME, Broggio J, Scott S, Loveday C, et al. Effect of delays in the 2-week-wait cancer referral pathway during the COVID-19 pandemic on cancer survival in the UK: a modelling study. Lancet Oncol. 2020 Aug;21(8):1035–44.

19. Al-Quteimat OM, Amer AM. The Impact of the COVID-19 Pandemic on Cancer Patients. Am J Clin Oncol. 2020 Jun;43(6):452–5.

20. Guckenberger M, Belka C, Bezjak A, Bradley J, Daly ME, DeRuysscher D, et al. Practice recommendations for lung cancer radiotherapy during the COVID-19 pandemic: An ESTRO-

ASTRO consensus statement. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2020 May;146:223–9.

21. van de Haar J, Hoes LR, Coles CE, Seamon K, Fröhling S, Jäger D, et al. Caring for patients with cancer in the COVID-19 era. Nat Med. 2020 May;26(5):665–71.

22. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet (London, England). 2020 Mar;395(10229):1054–62.

23. Sran K, Olsburgh J, Kasimatis T, Clark K, Gökmen R, Hilton R, et al. Coronavirus Disease 2019 in Kidney Transplant Patients From a Large UK Transplant Center: Exploring Risk Factors for Disease Severity. Transplant Proc. 2020 Dec;

24. Zimmerman A, Kalra D. Usefulness of machine learning in COVID-19 for the detection and prognosis of cardiovascular complications. Rev Cardiovasc Med. 2020 Sep;21(3):345–52.

Figure Legends

Figure 1: Relative importance of each variable in the machine learning algorithm in determining outcome from COVID-19 infection.

Figure 2: Generalized linear model identifying variables predicting mortality at 90 days from COVID-19.

Figure 3: Kaplan-Meier survival analysis and log-rank test to determine overall survival in cancer and non-cancer patients who contracted COVID-19.

Acknowledgement

The staff of Research and Development department of the Dudley Group Foundation Trust, West Midland for their support in conducting this study.

Author contributions:

Akinfemi Akingboye (MD, FRCS): Conceptualization, protocol development, ethics approval, proof-reading, manuscript writing, and approval of final manuscript

Fahad Mahmood (PhD, MRCS): Data analysis, Manuscript drafting and proof-reading

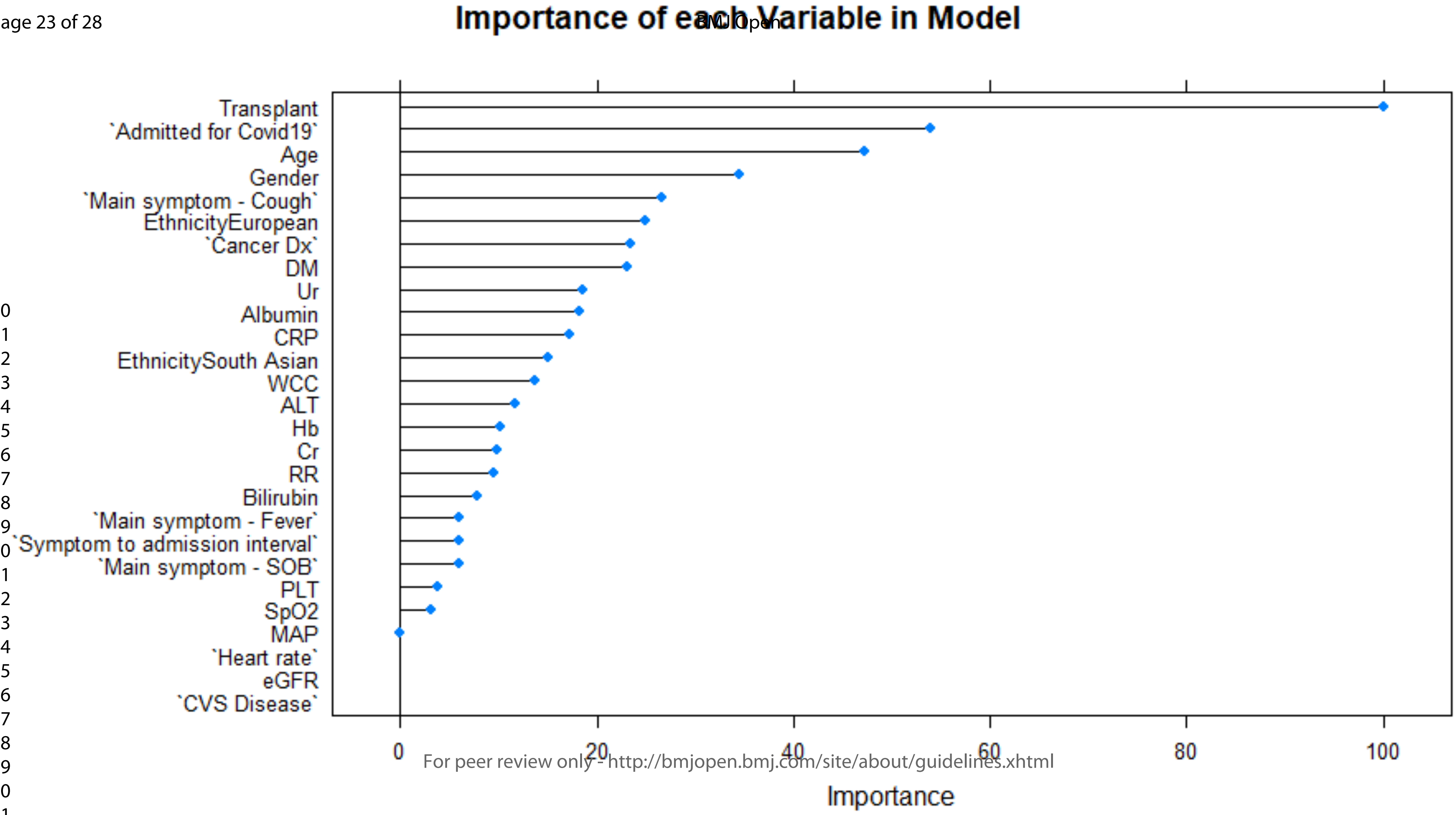
Nabeel Amiruddin (FRCA): Data analysis, machine learning algorithm and figures.

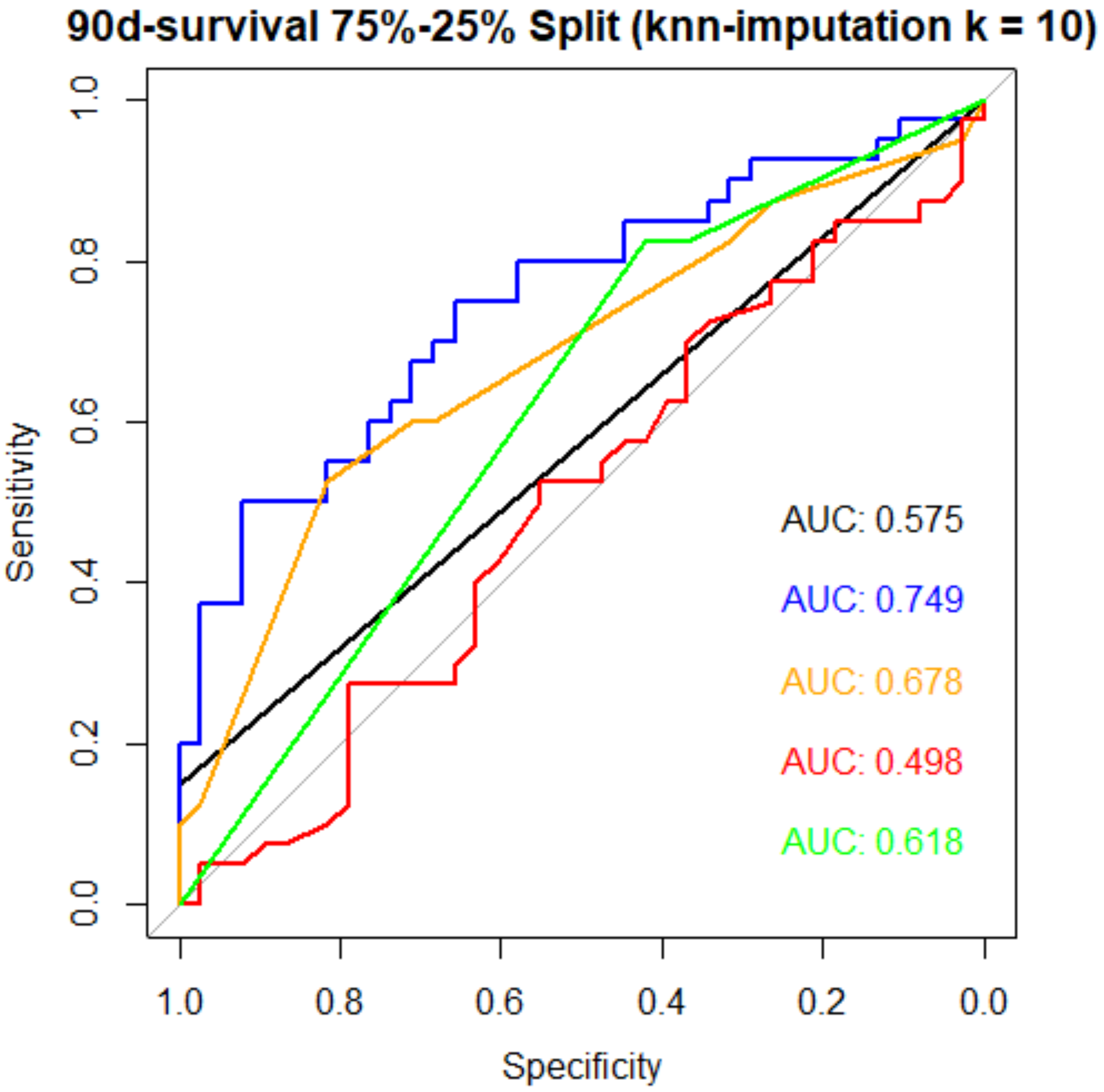
Michael Reay (FRCA): Conceptualization, data collection.

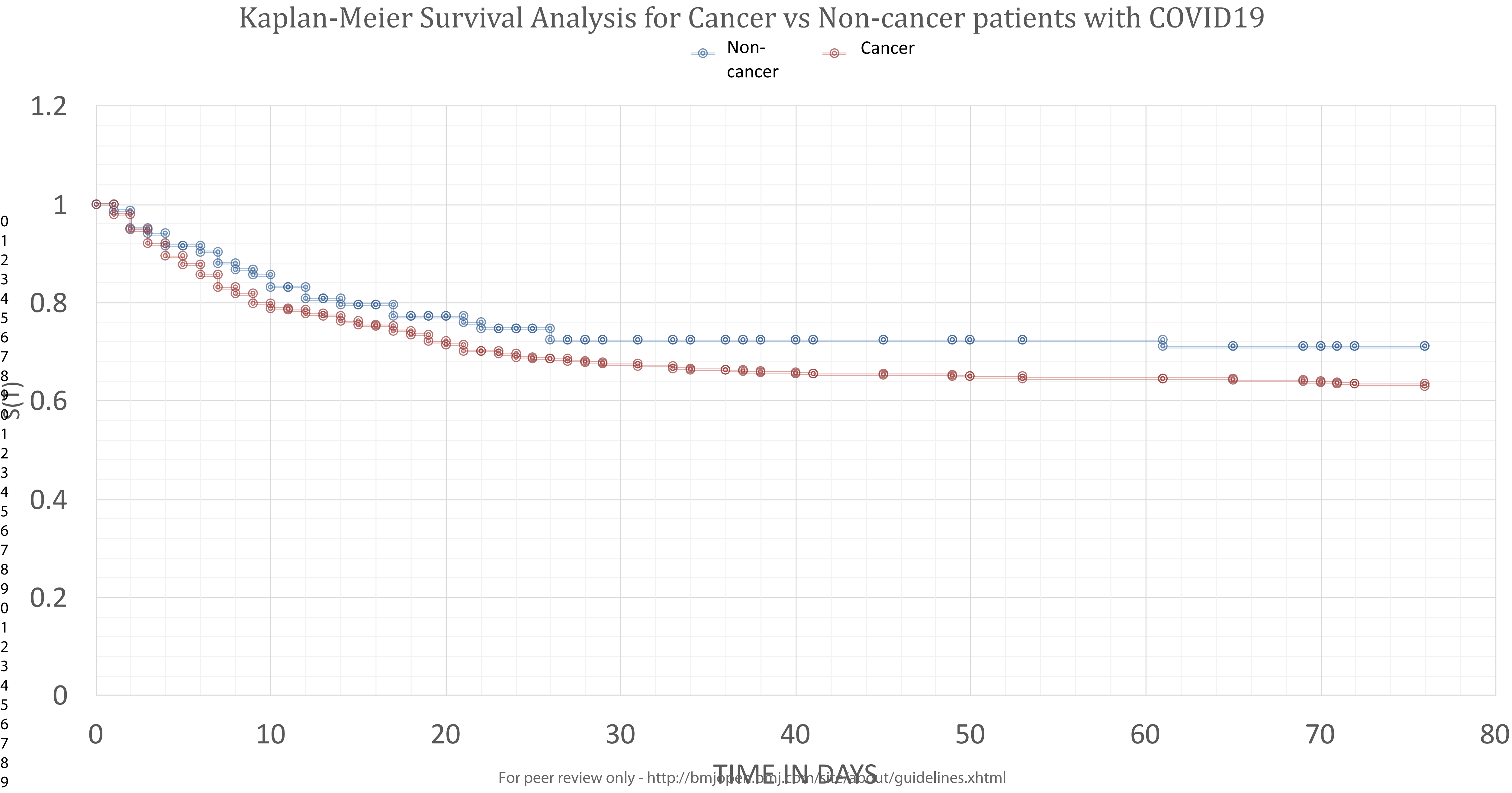
Peter Nightingale (PhD): Conventional statistical analysis.

Olorunseun Ogunwobi (MBBS, MSc, PhD): Conceptualization, reading and correcting manuscript drafts, and approval of final manuscript.

Patients or the public WERE NOT involved in the design, or conduct, or reporting, or dissemination plans of our research.

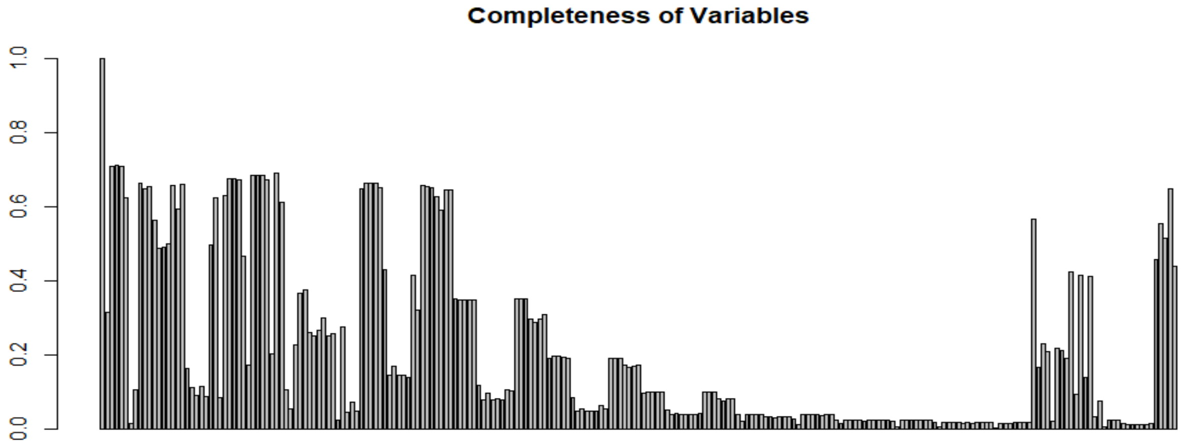

Importance of each Variable in Model
Transplant
`Admitted for Covid19`
Age
Gender
`Main symptom - Cough`
EthnicityEuropean
`Cancer Dx`
DM
Ur
Albumin
CRP
EthnicitySouth Asian
WCC
ALT
Hb
Cr
RR
Bilirubin
`Main symptom - Fever`
`Symptom to admission interval`
`Main symptom - SOB`
PLT
SpO2
MAP
`Heart rate`
eGFR
`CVS Disease`
0
20
40
60
80
100
Importance


90d-survival 75%-25% Split (knn-imputation k = 10)
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.575
AUC: 0.749
AUC: 0.678
AUC: 0.498
AUC: 0.618


Kaplan-Meier Survival Analysis for Cancer vs Non-cancer patients with COVID19
Non-cancer
Cancer
S(T)
1.2
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
TIME IN DAYS

Supplementary Figure 1

The graph below plots the proportion of missing data for each variable:



If you only include variables with at least 60% completeness, there are 33 variables out of 233.

Below is a feature plot of each these variables against whether or not the patient was alive at 30 days (the only dependent variable with >= 60% completeness):

Table 2: Univariate analysis of variables correlating with mortality risk in patients presenting with COVID19. Age, Hb, CRP, urea, creatinine, eGFR, low initial BP were associated with the highest risk.

	Univariable			95% CI
	p	OR	Lower	Upper
Age	0.000	1.048	1.031	1.064
White cell count (WCC)...57	0.407	1.005	0.994	1.016
Hemoglobin (Hb)...58	0.000	0.982	0.972	0.992
Platelets (Plt)...59	0.640	1.000	0.998	1.001
C Reactive Protein (CRP)...60	0.000	1.005	1.003	1.008
Urea (Ur)...69	0.000	1.132	1.081	1.185
Creatine (Cr)...70	0.001	1.006	1.003	1.010
Estimated glomoreluar filtration rate (eGFR)...71	0.000	0.977	0.967	0.988
Alanine transaaminase (ALT)...72	0.377	1.002	0.997	1.007
Bilirubin (Bili)...74	0.019	1.034	1.005	1.062

Albumin (Alb)...75	0.533	0.993	0.973	1.014
Initial vital signs – Heart rate	0.567	0.997	0.989	1.006
Initial vital signs – BP, Mean Arterial Pressure	0.048	0.987	0.973	1.000
Initial vital signs - RR	0.143	1.015	0.995	1.036
Initial vital signs - SpO2	0.454	0.996	0.986	1.006
Sex	0.212	0.759	0.491	1.171
Ethnicity	0.994			
Smoking	0.647			
Cardiovascular disease	0.024	1.671	1.069	2.614
Diabetes Mellitus	0.253	1.321	0.819	2.130
Transplant patient	0.166	4.740	0.524	42.877
Cancer patient	0.024	1.832	1.082	3.101
Reason for admission	0.143	0.610	0.315	1.181
Main symptom on admission - Cough	0.006	0.543	0.351	0.840
Main symptom on admission – fever of 38C and above	0.039	0.631	0.408	0.976
Main symptom on admission – Shortness of Breath	0.883	1.037	0.642	1.673

STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cohort studies*

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2-3
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	4
Methods			
Study design	4	Present key elements of study design early in the paper	5
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	5-8
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	5-8
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed	5-8
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	5-8
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	*5-8*
Bias	9	Describe any efforts to address potential sources of bias	5-8
Study size	10	Explain how the study size was arrived at	Not done
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	5-8
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	7-8
		(*b*) Describe any methods used to examine subgroups and interactions	7-8
		(*c*) Explain how missing data were addressed	7-8
		(*d*) If applicable, explain how loss to follow-up was addressed	7-8
		(*e*) Describe any sensitivity analyses	7-8
Results			

Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	8
		(b) Give reasons for non-participation at each stage	8-9
		(c) Consider use of a flow diagram	
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	8-9
		(b) Indicate number of participants with missing data for each variable of interest	8-9
		(c) Summarise follow-up time (eg, average and total amount)	8-9
Outcome data	15*	Report numbers of outcome events or summary measures over time	8-9
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	8-9
		(*b*) Report category boundaries when continuous variables were categorized	8-9
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	8-9
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	8-9
Discussion			10-13
Key results	18	Summarise key results with reference to study objectives	10-13
Limitations			
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	10-13
Generalisability	21	Discuss the generalisability (external validity) of the study results	10-13
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	3

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.

BMJ Open

Increased risk of COVID-19 related admissions in active solid organ cancer patients in the West Midlands region of the United Kingdom: A Retrospective cohort study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-053352.R1
Article Type:	Original research
Date Submitted by the Author:	29-Oct-2021
Complete List of Authors:	Akingboye, Akinfemi; Russells Hall Hospital Mahmood, Fahad; Russells Hall Hospital Amiruddin, Nabeel; Russells Hall Hospital Reay, Michael; Russells Hall Hospital Nightingale, Peter; University Hospitals Birmingham NHS Foundation Trust, Institute of Translational Medicine Ogunwobi, Olorunseun; Hunter College CUNY, Biological Sciences
<b>Primary Subject Heading</b>:	Oncology
Secondary Subject Heading:	Infectious diseases
Keywords:	COVID-19, Risk management < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, ONCOLOGY

Title page

Title: Increased risk of COVID-19 related admissions in active solid organ cancer patients in the West Midlands region of the United Kingdom: A Retrospective cohort study

Authors:

Akinfemi Akingboye (MD, FRCS)[1], Fahad Mahmood (PhD, MRCS)[1], Nabeel Amiruddin (FRCA)[1], Michael Reay (FRCA)[1], Peter Nightingale[2], Olorunseun Ogunwobi (MBBS, MSc, PhD)[3,4]

1. Russells Hall Hospital, Dudley, United Kingdom, DY1 2HQ
2. Institute of Translational Medicine, WCL - University Hospitals Birmingham NHS Foundation Trust
3. Hunter College Center for Cancer Health Disparities Research (CCHDR), New York
4. Hunter College of The City University of New York

Corresponding authors:

Olorunseun Ogunwobi, MBBS, MSc, PhD

Director, Hunter College Center for Cancer Health Disparities Research (CCHDR)

Associate Professor of Biological Sciences

Hunter College of The City University of New York

Belfer Research Building, Room 426, 413 E 69th Street, New York, NY 10021

Tel: 212-896-0447

E-mails: oo158@hunter.cuny.edu

Mr Akinfemi Akingboye MBBS, MD (Lond), FRCS (Gen. Surg.)

Consultant Laparoscopic Colorectal & General Surgeon

Department of General Surgery

The Dudley Group NHS Trust,

Russells Hall Hospital Dudley, DY1 2HQ

Email: a.akingboye@nhs.net

Tele: 01384 456111 Ext 2739

Word count: 3724

Abstract

Objective: Susceptibility of cancer patients to COVID-19 pneumonitis has been variable. We aim to quantify the risk of hospitalization in active cancer patients and use a machine learning algorithm (MLA) and traditional statistics to predict clinical outcomes and mortality.

Design: Retrospective cohort study.

Setting: A single United Kingdom (UK) district general hospital (Rec-Reference 20/EE/0139;IRAS ID28233).

Participants: Data on total hospital admissions between March 2018-June 2020, all active cancer diagnoses between March 2019-June 2020 and clinical parameters of COVID-19 positive admissions between March 2020 and June 2020 were collected. 526 COVID-19 admissions without an active cancer diagnosis were compared with 87 COVID-19 admissions with an active cancer diagnosis.

Primary & secondary outcome measures: 30 and 90-day post-COVID-19 survival.

Results: In total 613 patients were enrolled with Male to Female ratio was 1:6 and median age was 77 years. The estimated infection rate of COVID-19 was 87/22729 (0.4%) in the cancer patients and 526/404379 (0.1%) in the non-cancer population (OR of being hospitalised with COVID if having cancer is 2.942671 (95% CI: 2.344522-3.693425); p<0.001). Survival was reduced in cancer patients with COVID-19 at 90 days. R-Studio software determined the association between cancer status, COVID-19 and 90-day survival against variables using MLA. Multivariate analysis showed increases in age (OR1.039[95%CI1.020-1.057], p<0.001), urea (OR1.005[95%CI1.002-1.007],p<0.001) and CRP (OR1.065[95%CI1.016-1.116],p<0.008) is associated with greater 30-and 90-day mortality. The MLA model examined the contribution of predictive variables for 90-day survival (AUC: 0.749); with transplant patients, age, male gender and diabetes mellitus being predictors of greater mortality.

Conclusions: Active cancer diagnosis has a 3-fold increase in risk of hospitalization with COVID-19. Increased age, urea, and CRP predict mortality in cancer patients. MLA complements traditional statistical analysis in identifying prognostic variables for outcomes of COVID-19 infection in cancer patients. This study provides proof of concept for MLA in risk prediction for COVID19 in cancer patients and should inform a redesign of cancer services to ensure safe delivery of cancer care.

Trial Registration: Not applicable. Observational study, not a clinical trial.

Strength & Limitations of this Study

- The study uses novel analytical methods derived from machine learning to evaluate risk from COVID-19 in cancer patients from hospitalization to mortality.
- Statistical and machine learning methods are compared to develop a profile of factors that can worsen outcomes from COVID-19 in cancer patients.
- The study analyses COVID-19 outcomes in solid organ cancer patients in a cohort covering a single UK metropolitan region only. No haematological malignancies analysed.
- Patients with COVID-19 and cancer who did not require admission to hospital were not included in this study.

Funding Statement: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. However, Olorunseun O. Ogunwobi is supported by National Cancer Institute grant # CA221704.

Competing interests statement: No conflicts of interest to declare.

Introduction

The severe acute respiratory syndrome coronavirus 2 leads to the coronavirus disease 2019 (COVID-19) (1,2). This highly transmissible disease has led to a global pandemic contributing to significant morbidity and mortality. Increased susceptibility and severity of COVID-19 are attributed to increasing age, smoking status, chronic obstructive pulmonary disease, diabetes mellitus, obesity, cardiovascular disease as well as cancer (3–6). In addition, the prevalence of all types of active or previous cancer in the United Kingdom (UK) is reported at 2.5 million cases with an incidence of 1000 newly diagnosed cases each day (7). Increased susceptibility to COVID-19 in cancer patients has been attributed to immune suppression and cancer treatments such as cytotoxic chemotherapy and immunotherapy (8,9). However, it is still not established whether this translates into increased hospitalization, illness severity or mortality risk. Risk adjusted models quote a mortality risk of between 25-39% in cancer patients hospitalized with COVID-19 (10). With increasing prevalence of COVID-19 in the UK, the impact of cancer on COVID-19 remains an area of active concern (8).

In addition, machine learning algorithms (MLA) have become increasingly applied in healthcare settings due to their prognostic utility (11–13). They are able to map a large number of observed variables (features) to target outcomes, and through statistical analysis, find relationships without human instruction (14,15). This utility has been exploited in cancer research to model risk of susceptibility, survival and recurrence (15). For example, in breast cancer, algorithms have been developed from detecting breast tumours to determining the prognostic significance of the tumour's morphological features (14). Moreover, the ability to integrate diverse variables including clinical, biochemical, histopathological, genomic and proteomic data could lead to more reliable predictive models to determine disease outcome (16,17). Furthermore, the scalability of MLA distinguishes it from traditional statistical modelling, such as regression analysis, by its ability to perform non-linear modelling utilizing large volume datasets and greater number of variables from registries (11,12). MLA models for risk prediction are starting to be validated in large studies (18).

Thus, this nascent technique holds promise for developing better risk assessment and prognostic algorithms to support healthcare delivery and individualized patient care.

This study aims to quantify the risk of hospitalization in active cancer patients using specific differences in clinicopathological and biochemical parameters between cancer and non-cancer COVID-19 patients through developing a machine learning algorithm. We seek to identify the most important determinants of high risk of susceptibility and mortality from a diverse range of variables. This will both provide proof of concept for our method as well as inform the recalibration of cancer services to ensure safe care for cancer patients during the pandemic.

Methods

A single UK centre retrospective cohort study was conducted **(Rec-Reference 20/EE/0139; IRAS ID28233).** Ethical approval was obtained from the Russells Hall ethics committee. Data on total hospital admissions between March 2018 and June 2020 was obtained from the local Information Technology (IT) department with a record of all hospital admission before and during the COVID-19 pandemic linking this with the Somerset cancer database to extrapolate the total number of active cancer patients that were admitted during the study period. Furthermore, all active solid organ cancer diagnoses between March 2019 and June 2020 were obtained from the local cancer network. This was used to determine the total number of active cancer patients with COVID-19 with the denominator being the total and active cancer population in the Dudley, West Midlands (UK) region. Biochemical and haematological parameters in the first 48 hours of admission along with 30 and 90-day post-COVID-19 survival was determined.

Patients below the age of 18 years and those with non-solid organ cancers were excluded. Moreover, patients who attended the emergency department and were not admitted were also excluded. COVID-19 diagnosis was established with a positive Reverse Transcriptase Polymerase Chain reaction (RT-PCR) test from an oropharyngeal swab Criteria for admission to hospital and critical care were determined by individual clinical assessment and oxygen requirement as well as ventilatory support. Data security was maintained through the RedCap uploading system.

Binary logistic regression analyses with survival status at 30 days as the dependent variable were used to estimate the univariable association with mortality for each explanatory variable. Age-adjusted associations were calculated in a similar way by including age at admission as a continuous variable in each model after checking the assumption of a linear effect of age on the log odds. Both forward and backward stepwise methods were used to determine the final multivariable model. These analyses were performed with SPSS 25.0.

Patient and public involvement

Patients and the public were not involved in the design and conduct of this study.

Machine Learning Algorithm: Data pre-processing

R-Studio software was used to determine the association between cancer status, COVID-19 and 90-day survival against variables in a Machine Learning Algorithm. The conduct and reporting of our MLA was done in accordance with best practice guidance (11).

Feature selection:

The proportion of missing data was calculated for each variable and variables with less than 40% missing data were included in the analysis. This resulted in 33 variables being included for imputation of missing data, further pre-processing and model development. The decision to limit the proportion of missing data to 40% was an arbitrary one, based on a compromise between a limit high enough to enable the inclusion of as many available variables as possible and low enough to enable the use of more data to predict imputable missing values with the k-nearest neighbours' algorithm.

All data within the Gender variable was replaced with 'F'(female) and 'M'(male). The documented ethnicities were replaced with the 3 categories of 'European', 'South-Asian' or 'Afro-Caribbean'. The blood pressure information was split into systolic and diastolic pressures. A new dummy variable of 'Mean arterial pressure' was derived from the estimate: (Diastolic pressure + (Pulse pressure/3)). The pure numeric values from the entered data for oxygen saturations were extracted. For example, 97% would be changed to 97. A dummy variable was created from the difference in time between the date of onset of symptoms and date of hospital admission. This time-interval was recorded in days.

Data partition:

The overall dataset was partitioned into training and test sets. The training set was used purely for model training and hyperparameter tuning. The test set would only be used for model evaluation against new data. Partitioning was by a random allocation, while ensuring an identical distribution of patients who died at 90 days between both training and test sets. 75% of patients were allocated to the training set, with the remaining going into the test set (Supplementary Figure 1).

Imputation of missing values:

Missing values were replaced with predicted values using k-nearest neighbour's model. This method designated the variable of a missing value as an outcome variable within a predictive model. A prediction of the missing value on the most similar k number of patients based on their other variables. The value k is a hyperparameter which was set to 10 after comparing the values of 5 and 10 without any difference.

This imputation was performed separately on the training and test datasets in order to minimise overfitting of the final model by having the training dataset influence the imputation of values into the test dataset.

Additional pre-processing:

All numeric variables within the training set were pre-processed for model training to be on comparable scales ranging mainly from 0 to 1. For each such variable, the mean was subtracted from each value before the dividing the result by the standard deviation.

The same process was applied to the test set, using the means and standard deviations from the training set to avoid overfitting.

Model development:

The following models were trained using 10-fold cross validation:

1. Logistic regression

2. Lasso and Elastic-net generalised linear model
3. K-nearest neighbour
4. Random forest
5. Neural network with 1 hidden layer
6. Gradient boosted machine

Hyperparameter tuning during cross validation was optimised against area under the receiver operator curve as a metric. The random forest model was built with 500 trees.

Model evaluation:

Predictions of probabilities of survival to 90 days was made on the test set by each of the 5 trained models. The known survival outcomes to 90 days and predicted probabilities from each model were used to plot receiver-operator curves for model for comparison.

Results

In total, 22,729 active cancer patients were identified in the Dudley West Midlands region out of a catchment size of 426658 patients in the region from the local cancer network. 87/22729 (0.4%) cancer patients in the Dudley region were admitted with COVID-19 compared with 526/404379 (0.1%) during the study period (Hazard ratio: OR of being hospitalised with COVID if having cancer is 2.942671 (95% CI: 2.344522-3.693425); p<0.001). The types of cancer in our cohort are detailed in Figure 1. Thus, the risk of hospital admission on presentation with COVID-19 increased 3-fold in the presence of an active cancer diagnosis.

Excluding those with incomplete data, the mean age of cancer patients was 77.8(sd=12.3) years compared to 70 (sd=17.5) years (t-test; p<0.001). The Male:Female ratio was similar between the two groups. The majority of patients were of Caucasian ethnicity with similar distribution of diabetes, cardiovascular disease, transplant recipient and smoking status. Moreover, the median white cell count (p=0.096) and CRP (p=0.115) were similar between cancer and non-cancer patients with no statistically significant difference. Thus, both cancer and non-cancer groups affected by COVID-19 had similar baseline characteristics. This is summarized in Table 1.

Table 1: Comparing characteristics of cancer and non-cancer patients. Values are counts and percentages except where stated. The p values are from Fisher's exact test, except for age (from a t test), white cell count and CRP (both from Mann-Whitney tests).

	CANCER PATIENTS (N=80)	NON-CANCER PATIENTS (N=276)	P VALUE
AGE IN YEARS (N=356) : MEAN (SD)	77.8 (12.3)	70.0 (17.5)	<0.001
SEX			0.699
FEMALE	34 (43%)	112 (41%)	
MALE	45 (57%)	164 (59%)	
ETHNICITY			0.280
AFRO-CARIBBEAN	1 (1%)	9 (4%)	

EUROPEAN	70 (95%)	197 (87%)	
SOUTH ASIAN	3 (4%)	50 (9%)	
SMOKING			0.176
CURRENT	5 (22%)	18 (30%)	
EX	9 (39%)	11 (18%)	
NEVER	9 (39%)	31 (52%)	
CARDIOVASCULAR			0.103
YES	22 (31%)	107 (41%)	
NO	50 (69%)	152 (59%)	
DIABETES MELLITUS			0.885
YES	22 (29%)	73 (28%)	
NO	53 (71%)	187 (72%)	
TRANSPLANT PATIENT			0.644
YES	2 (3%)	5 (2%)	
NO	70 (97%)	261 (98%)	
REASON FOR ADMISSION			<0.001
YES	48 (70%)	227 (90%)	
NO	21 (30%)	24 (10%)	
WHITE CELL COUNT (N=332) : MEDIAN (LOWER QUARTILE – UPPER QUARTILE)	8.8 (5.6-12.7)x10^9/L	7.2 (5.3-10.6)x10^9/L	0.096
CRP (N=324) : MEDIAN (LOWER QUARTILE – UPPER QUARTILE)	77 (22-135) mg/L	84 (36-157) mg/L	0.115

A chi-squared test, comparing non-cancer patients not hospitalised with COVID-19 (404379) yields a p-value of <2.2e-16, implying that there is an association between having cancer and hospitalisation with COVID-19 (Table 2). The Odds ratio of being hospitalised with COVID-19 if having cancer is 2.942671 (95% CI: 2.344522-3.693425).

Table 2: Risk of admission with COVID-19 in cancer patients. Pearson's Chi-squared test with Yates' continuity correction: X-squared = 93.641, df = 1, p-value < 2.2e-16.

Group	Admitted	Not Admitted
Non- cancer patients	404379	526
Cancer patients	87	22729

After training and hyperparameter tuning by 10-fold cross-validation, predictions of probability of 90-day survival were made on the test-set data. This is shown in the receiver-operator curves plotted for model comparison (Figure 2).

The random forest model achieved an area under the receiver-operator curve (AUROC) of 0.829 (Figure 2). Each variable was evaluated for its relative contribution to enabling classification (Figure 3).

Since we accepted variables with up to 40% missing values (Supplementary Figure 2), imputation was performed using a separate k-nearest neighbours' algorithm, whereby a prediction of a missing value was made based the other available values, having been trained on the other patient data.

Our initial age-adjusted univariate analysis identified age, CRP, urea, creatinine, eGFR, haemoglobin and low initial blood pressure as significantly correlating with mortality risk (Supplementary Table 1). A further multivariate analysis of 33 out of 213 clinical variables with >60% data completeness showed increased age (Hazard ratio 0.915 [95%CI 0.870-0.960], p<0.001), urea (Hazard ratio 1.005 [95%CI 1.002-1.007], p<0.001) and CRP (Hazard ratio 1.065 [95%CI1.016-1.116], p<0.001) to be associated with greater risk of 30-and 90-day mortality (Table 3).

Table 3: Multivariate analysis showing increased age, CRP and urea are associated with the highest 90-day mortality risk in COVID-19 patients.

	P-valve	*OR*	*Lower CI*	*Upper CI*
Age	0.000	1.039	1.020	1.057
CRP	0.001	1.005	1.002	1.007
Urea	0.008	1.065	1.016	1.116

Kaplan-Meier survival analysis revealed reduced overall survival for patients with COVID-19 and cancer (Figure 4). However, Log-Rank analysis did not show significant difference between cancer and non-cancer COVID-19 patients (Log-rank p=0.172).

Discussion

Our study demonstrates that the presence of active cancer increased by 3-fold the risk of hospitalization with COVID-19. Moreover, higher CRP and urea are associated with greater mortality at 30- and 90-days post-diagnosis of COVID-19. These findings show that cancer patients who develop COVID-19 are likely to have a more severe form of the infection that would require supportive care in hospital. It also provides tools for monitoring patient response to treatment with high urea and CRP being poor prognostic markers and a likely consequence of severe COVID-19. This has implications for how we can deliver safe care to cancer patients in the ongoing pandemic as well as emerging from it given the restrictions on cancer services.

Several studies have reported prevalence and mortality risk of COVID-19 in cancer patients with a systematic review by Zarifkar *et al* identifying 110 studies covering 10 countries (19). The pooled prevalence of active cancer in COVID-19 positive hospitalized patients was 2.6% (95% CI 1.8-3.5%) across 37 cohort studies. Furthermore, there was a noticeable difference in the prevalence between western countries (5.6%, 95% CI 4.5%-6.7%) and China (1.7%, 95% CI 1.3%-2.3%) reflecting the underlying cancer prevalence. In addition, in-hospital mortality of 14.1% (95% CI 9.1-19.8%) for cancer and COVID-19 was derived from 17 retrospective cohort studies covering 904 patients (19). The mortality rate of 12.6% in a Brazilian cohort was also similarly reported (20). This indicated that COVID-19 patients with cancer had a 5-fold greater risk of death compared with non-cancer patients without other co-morbidities (19,21). However, there was significant heterogeneity between these studies (I^2=55.9%, P<0.01) with the type of cancer, stage and treatment regimen only specified in 8 studies along with incomplete followup. Furthermore, Liang et al reported a 28% prevalence of lung cancer amongst hospitalized cancer patients with COVID-19 (9). This reflects higher COVID-19 mortality rates in specific cancer patients including lung and haematological malignancy (9,19,20). Further studies have reported 3.5 fold increase in ICU admission or need for mechanical ventilation in COVID-19 patients with cancer (9). More recent studies have also examined the impact of COVID-19 on cancer patients. In an analysis of 306 COVID-19 patients with cancer, Russell et al identified factors including male gender, age greater than 60, Asian ethnicity, cancer diagnosis of greater than 2 years, haematological malignancy and a high CRP associated with increased mortality risk (22). A large population based study by Lee et al comparing 23,266 cancer patients with 1784293 non-cancer patients identified a 60% increased risk of COVID-19 in cancer patients with those on chemotherapy or immunotherapy having a 2.2-fold increased risk of contracting COVID-19 (23). This increased susceptibility could be explained through immune compromise of simply greater exposure through more frequent hospital visits. Even in this large study, subgroup analysis was not performed evaluating the impact of tumour type, stage and treatment regimens. Furthermore, in a multicentre study comparing UK (n=468) and EU (n=924) cancer patients with COVID-19, showed a worse mortality rate at 30 days and 6 months independent of age, gender, tumour stage and treatment through a multivariable regression model (24). Moreover, Mehta et al showed increased risk of COVID-19 in 218 cancer patients in New York which was associated with age, co-morbidities and elevated lactate dehydrogenase (25). Although few studies have indicated no increased risk of severity or mortality from COVID-19 in cancer patients (26,27), the consensus thus far in the literature has coalesced around the idea that cancer patients in general have a higher risk of susceptibility from severe events and mortality from COVID-19 infection. Using admission risk as a surrogate marker of severity in COVID-19, our results are consistent with the literature showing a 3-fold higher risk of admission with COVID-19 in the presence of cancer which will likely impact the delivery of care to these particular subgroups of patients. However, the majority of current data is from retrospective cohort studies, using traditional statistical techniques on selected limited variables with a relatively small number of participants. Moreover, for this reason subgroup analysis has been difficult. Since cancer is a diverse condition from the clinical to genomic spheres, with an equally diverse range of treatments, considering it as a monolithic structure would not let meaningful conclusions to be drawn from such analyses. Having a multicentre approach and application of novel big-data analysis techniques such as MLA may enable a more reliable and rapid analysis of data to discover associations in time-critical situations such as delivering healthcare in a global pandemic.

There is likely to be a surge in demand for cancer services as well as predicted poor long-term survival in cancer patients due to delays in diagnosis and treatment (11,28). Over the first national UK lockdown, there was a 84% reduction in urgent cancer referrals which modelling predicted would lead to 181 additional lives lost or 3316 life-years lost with an average presentation delay of 2 months per patient (29). Although having cancer puts patient at increased risk of hospitalization with COVID-19, this must be balanced against risks of delayed treatment leading to disease progression to incurable stages (30). Particular cancers where timely intervention is critical such as pancreatic, lung and haematological malignancy should not have delays to treatment whereas others including prostate and non-melanoma skin cancers treatment may be safely delayed in selected patients (30). Several strategies including: Delays to surgery or chemotherapy, switching to oral or monotherapy treatment regimens, strict infection control protocols, online consultation, use of hypofractionated radiotherapy and provision of intensive care support to these patients is essential to mitigate risk (10,30–32). This may be supplemented where possible with COVID-19 free 'cold' sites to reduce risk of transmission and prevent anti-cancer treatment induced COVID-19 (32). Thus categorization of patients according to risk, minimizing patient exposure and considering alternative regimens to control cancer forms the basis of current recommendations including the ESMO expert consensus and UK NHS guidelines (4,32). Furthermore, this data does not support delays in cancer treatment to reduce risk of COVID-19 transmission in cancer patients.

Several biochemical markers have been associated with a severe COVID-19 disease course. Zhou *et al* identified a raised D-dimer above 1μg/ml to be associated with a higher mortality risk (33). Furthermore, they identified low albumin, raised LDH, troponin, ferritin and IL-6 to be more prevalent in non-survivors. In addition, raised CRP and low glomerular filtration rate was associated with a more severe disease outcome with 18% deaths recorded in a renal transplant cohort in keeping with the severe disease course predicted in this group of immunocompromised patients (34).Our model identified raised urea and CRP in addition to transplant status as predictors of greater mortality risk which may lower threshold for admission or earlier referral for intensive care support. However, our algorithm could not specify the direction or size of this interaction which is a limitation of such models.

Machine learning algorithms are increasingly being used to support healthcare applications including cancer diagnosis, outcomes and recurrence (16,17,35). MLA can be used to learn from established data sets and identify hidden patterns between a large number of variables to support individualized decision making (36). Nonetheless, the technique requires training datasets, appropriately selected analysis method, as well as testing datasets to establish internal and external validity (16,35). MLA have been shown to improve the accuracy of predicting cancer susceptibility, recurrence and mortality by 15-25% (37). Moreover, we have shown through our modelling that the findings of both MLA and traditional statistical analysis are complementary and may be used to generate a risk prediction scoring system in cancer patients with COVID-19.

However, there are several limitations in the method and data presented. MLA remains an experimental technique and still very dependent on the quality of input data. Issues including noise, bias, outliers, missing or duplicate data can lead to mis-classifications in any risk prediction model which may be mitigated by larger datasets (15). As such, few MLAs have achieved validation or widespread clinical application. In our study, confounders including smoking status or respiratory co-morbidity were not assessed which could influence outcomes in cancer patients. Patients with active cancer who tested positive for COVID-19 in the community but did not require hospital admission could not be evaluated. Having a broad inclusion criterion with all solid organ cancers whilst beneficial for looking at overall impact on cancer patients does not capture the granularity of how individual cancers may differ in their impact on COVID-19 patients. For example, haematological malignancy, lung cancer and metastatic disease were associated with adverse outcomes from COVID-19 infection (38). Our dataset was underpowered to perform relevant sub-group analyses on these patients. Although all active cancer patients were analysed, variation in the stage of cancer and treatment protocols were not accounted. Thus, the machine learning algorithms are limited by the quality of data input and rely upon imputation as part of model development which needs external validation once developed which we have not performed. Nonetheless this study provides proof-of-concept to investigate this question in a collaborative manner using larger datasets.

Conclusions

COVID-19 has impacted both individuals and healthcare systems in an enormous way. How we deliver safe and effective care to these patients in the confines of our healthcare systems is predicated on identifying those most as risk from this disease. Machine learning algorithms provide an additional tool to for risk assessment to delineate factors with poor prognosis. This will enable us to reconfigure our healthcare systems to provide safe care to these more vulnerable patients.

Data availability statement: All data relevant to the study are included in the article or uploaded as supplementary information.

Ethical Approval Statement: This study involves human participants and was approved by Russells Hall Ethics Committee (Rec-Reference 20/EE/0139; IRAS ID28233).

References

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet (London, England). 2020 Feb;395(10223):497–506.
2. Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. First Case of 2019 Novel Coronavirus in the United States. N Engl J Med. 2020 Mar;382(10):929–36.
3. Emami A, Javanmardi F, Pirbonyeh N, Akbari A. Prevalence of Underlying Diseases in Hospitalized Patients with COVID-19: a Systematic Review and Meta-Analysis. Arch Acad Emerg Med. 2020;8(1):e35.
4. Curigliano G, Banerjee S, Cervantes A, Garassino MC, Garrido P, Girard N, et al. Managing cancer patients during the COVID-19 pandemic: an ESMO multidisciplinary expert consensus. Vol. 31, Annals of oncology : official journal of the European Society for Medical Oncology. 2020. p. 1320–35.
5. Jordan RE, Adab P, Cheng KK. Covid-19: risk factors for severe disease and death. Vol. 368, BMJ (Clinical research ed.). England; 2020. p. m1198.
6. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA. 2020 May;323(20):2052–9.
7. MacMillan Trust. Macmillan Cancer Support Cancer in numbers. [Internet]. Macmillan Cancer Support. 2020. Available from: https://www.macmillan.org.uk/about-us/media-centre/facts-and-figures/cancer-in-numbers.html
8. UKCCM. The UK Coronavirus Cancer Monitoring Project: protecting patients with cancer in the era of COVID-19. Lancet Oncol. 2020 May;21(5):622–4.
9. Liang W, Guan W, Chen R, Wang W, Li J, Xu K, et al. Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China. Lancet Oncol. 2020 Mar;21(3):335–7.

10. Abdihamid O, Cai C, Kapesa L, Zeng S. The Landscape of COVID-19 in Cancer Patients: Prevalence, Impacts, and Recommendations. Cancer Manag Res. 2020;12:8923–33.

11. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016 Dec;18(12):e323.

12. Scott IA. Demystifying machine learning - a primer for physicians. Intern Med J. 2021 Jan;

13. Sarijaloo F, Park J, Zhong X, Wokhlu A. Predicting 90 day acute heart failure readmission and death using machine learning-supported decision analysis. Clin Cardiol. 2021 Feb;44(2):230–7.

14. Ibrahim A, Gamble P, Jaroensri R, Abdelsamea MM, Mermel CH, Chen P-HC, et al. Artificial intelligence in digital breast pathology: Techniques and applications. Breast. 2020 Feb;49:267–73.

15. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M V, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8–17.

16. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol. 2019 May;20(5):e262–73.

17. Radakovich N, Nagy M, Nazha A. Machine learning in haematological malignancies. Lancet Haematol. 2020 Jul;7(7):e541–50.

18. Ravaut M, Harish V, Sadeghi H, Leung KK, Volkovs M, Kornas K, et al. Development and Validation of a Machine Learning Model Using Administrative Health Data to Predict Onset of Type 2 Diabetes. JAMA Netw open. 2021 May;4(5):e2111315.

19. Zarifkar P, Kamath A, Robinson C, Morgulchik N, Shah SFH, Cheng TKM, et al. Clinical Characteristics and Outcomes in Patients with COVID-19 and Cancer: a Systematic Review and Meta-analysis. Clin Oncol (R Coll Radiol). 2021 Mar;33(3):e180–91.

20. Fernandes GA, Feriani D, França e Silva ILA, Mendonça e Silva DR, Arantes PE, Canteras J da S, et al. Differences in mortality of cancer patients with COVID-19 in a Brazilian cancer center. Seminars in Oncology. 2021.

21. Ioannidis JPA, Axfors C, Contopoulos-Ioannidis DG. Population-level COVID-19 mortality risk

for non-elderly individuals overall and for non-elderly individuals without underlying diseases in pandemic epicenters. Environ Res. 2020 Sep;188:109890.

22. Russell B, Moss CL, Shah V, Ko TK, Palmer K, Sylva R, et al. Risk of COVID-19 death in cancer patients: an analysis from Guy's Cancer Centre and King's College Hospital in London. Br J Cancer. 2021 Sep;125(7):939–47.

23. Lee KA, Ma W, Sikavi DR, Drew DA, Nguyen LH, Bowyer RCE, et al. Cancer and Risk of COVID-19 Through a General Community Survey. Oncologist. 2021 Jan;26(1):e182-5.

24. Pinato DJ, Scotti L, Gennari A, Colomba-Blameble E, Dolly S, Loizidou A, et al. Determinants of enhanced vulnerability to coronavirus disease 2019 in UK patients with cancer: a European study. Eur J Cancer. 2021 Jun;150:190–202.

25. Mehta V, Goel S, Kabarriti R, Cole D, Goldfinger M, Acuna-Villaorduna A, et al. Case Fatality Rate of Cancer Patients with COVID-19 in a New York Hospital System. Cancer Discov. 2020 Jul;10(7):935–41.

26. Liu C, Zhao Y, Okwan-Duodu D, Basho R, Cui X. COVID-19 in cancer patients: risk, clinical features, and management. Cancer Biol Med. 2020 Aug;17(3):519–27.

27. Vuagnat P, Frelaut M, Ramtohul T, Basse C, Diakite S, Noret A, et al. COVID-19 in breast cancer patients: a cohort at the Institut Curie hospitals in the Paris area. Breast Cancer Res. 2020 May;22(1):55.

28. Oncology TL. COVID-19: global consequences for oncology. Vol. 21, The Lancet. Oncology. 2020. p. 467.

29. Sud A, Torr B, Jones ME, Broggio J, Scott S, Loveday C, et al. Effect of delays in the 2-week-wait cancer referral pathway during the COVID-19 pandemic on cancer survival in the UK: a modelling study. Lancet Oncol. 2020 Aug;21(8):1035–44.

30. Al-Quteimat OM, Amer AM. The Impact of the COVID-19 Pandemic on Cancer Patients. Am J Clin Oncol. 2020 Jun;43(6):452–5.

31. Guckenberger M, Belka C, Bezjak A, Bradley J, Daly ME, DeRuysscher D, et al. Practice

recommendations for lung cancer radiotherapy during the COVID-19 pandemic: An ESTRO-ASTRO consensus statement. Radiother Oncol J Eur Soc Ther Radiol Oncol. 2020 May;146:223–9.

32. van de Haar J, Hoes LR, Coles CE, Seamon K, Fröhling S, Jäger D, et al. Caring for patients with cancer in the COVID-19 era. Nat Med. 2020 May;26(5):665–71.

33. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet (London, England). 2020 Mar;395(10229):1054–62.

34. Sran K, Olsburgh J, Kasimatis T, Clark K, Gökmen R, Hilton R, et al. Coronavirus Disease 2019 in Kidney Transplant Patients From a Large UK Transplant Center: Exploring Risk Factors for Disease Severity. Transplant Proc. 2020 Dec;

35. Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. Cancer Lett. 2020 Feb;471:61–71.

36. Zimmerman A, Kalra D. Usefulness of machine learning in COVID-19 for the detection and prognosis of cardiovascular complications. Rev Cardiovasc Med. 2020 Sep;21(3):345–52.

37. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2007 Feb;2:59–77.

38. Dai M, Liu D, Liu M, Zhou F, Li G, Chen Z, et al. Patients with Cancer Appear More Vulnerable to SARS-CoV-2: A Multicenter Study during the COVID-19 Outbreak. Cancer Discov. 2020 Jun;10(6):783–91.

Figure Legends

Figure 1: Types of active cancer in our cohort of patients for analysis. Solid organ and skin cancers were grouped together for analysis.

Figure 2: Receiver-operator curves for logistic regression (black), Generalised linear model (blue), k-nearest neighbours (orange), Random Forest (red), single hidden layer neural network (green), Gradient boosted machine (brown).

Figure 3: Relative importance of each variable in the machine learning algorithm in determining outcome from COVID-19 infection.

Figure 4: Kaplan-Meier survival analysis and log-rank test to determine overall survival in cancer and non-cancer patients who contracted COVID-19.

Supplementary Figure 1: Relative importance of each variable in the machine learning algorithm in determining outcome from COVID-19 infection.

Supplementary Figure 2: Receiver-operator curves for logistic regression (black), Generalised linear model (blue), k-nearest neighbours (orange), Random Forest (red), single hidden layer neural network (green), Gradient boosted machine (brown). 75% of patients were allocated to train the model.

Acknowledgement

The staff of Research and Development department of the Dudley Group Foundation Trust, West Midland for their support in conducting this study.

Author contributions:

Akinfemi Akingboye (MD, FRCS): Conceptualization, protocol development, ethics approval, proof-reading, manuscript writing, and approval of final manuscript

Fahad Mahmood (PhD, MRCS): Data analysis, Manuscript drafting and proof-reading

Nabeel Amiruddin (FRCA): Data analysis, machine learning algorithm and figures.

Michael Reay (FRCA): Conceptualization, data collection.

Peter Nightingale (PhD): Conventional statistical analysis.

Olorunseun Ogunwobi (MBBS, MSc, PhD): Conceptualization, reading and correcting manuscript drafts, and approval of final manuscript.

Patients or the public WERE NOT involved in the design, or conduct, or reporting, or dissemination plans of our research.

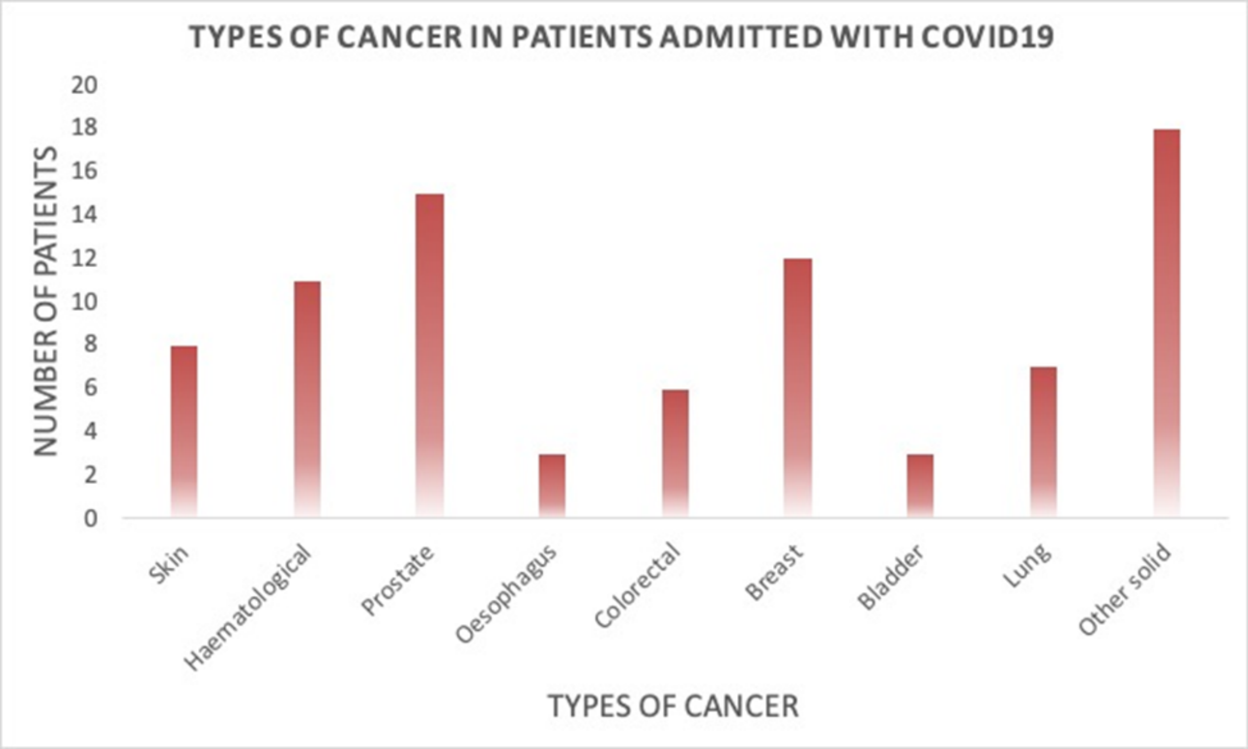


Figure 1: Types of active cancer in our cohort of patients for analysis. Solid organ and skin cancers were grouped together for analysis.

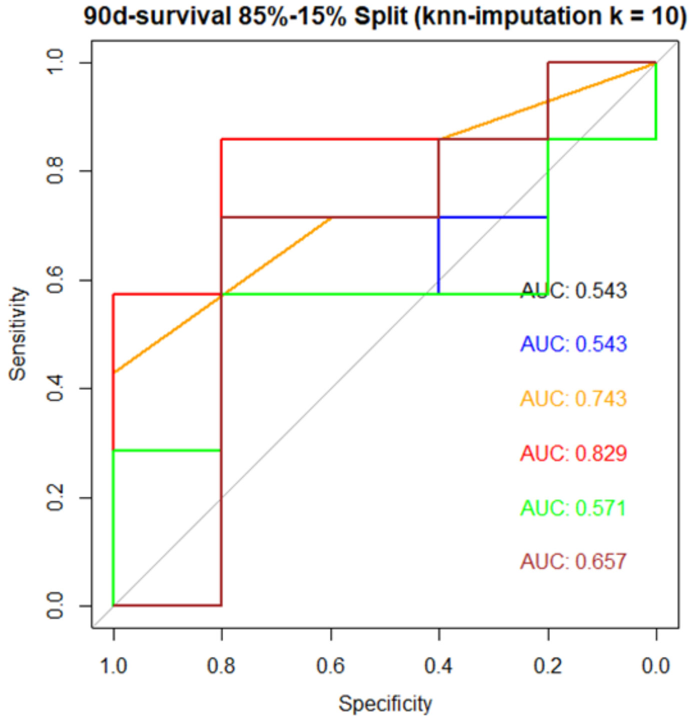


Figure 2: Receiver-operator curves for logistic regression (black), Generalised linear model (blue), k-nearest neighbours (orange), Random forest (red), single hidden layer neural network (green), Gradient boosted machine (brown)

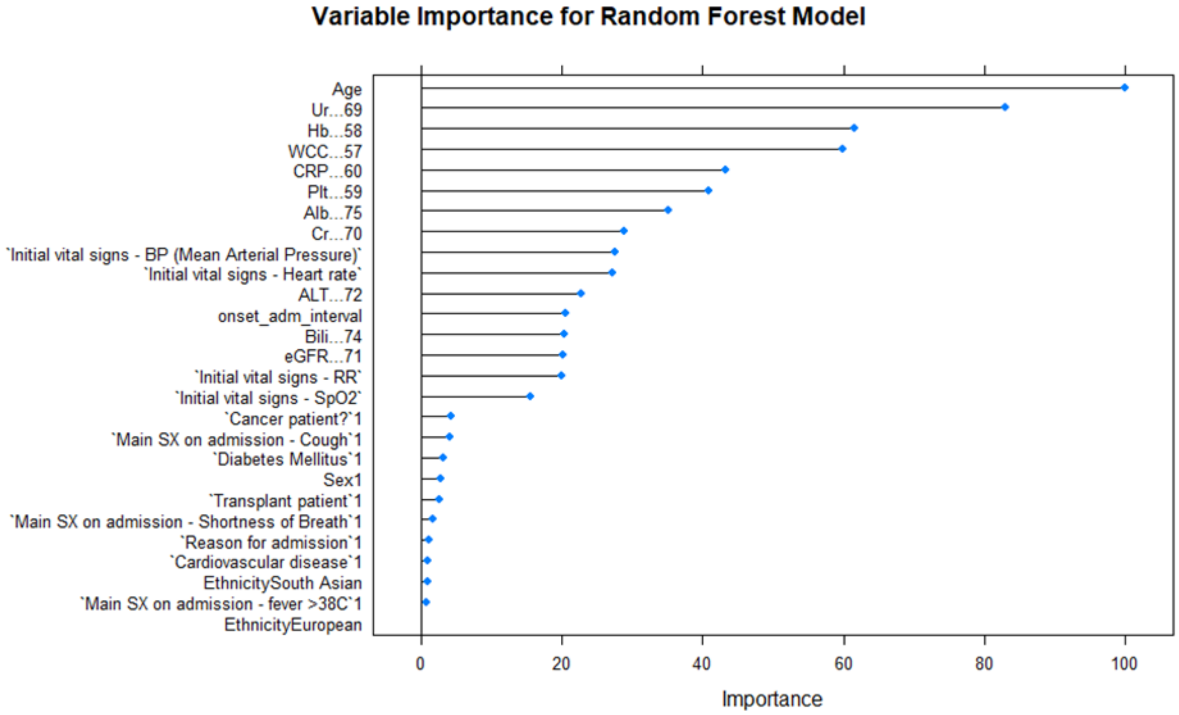
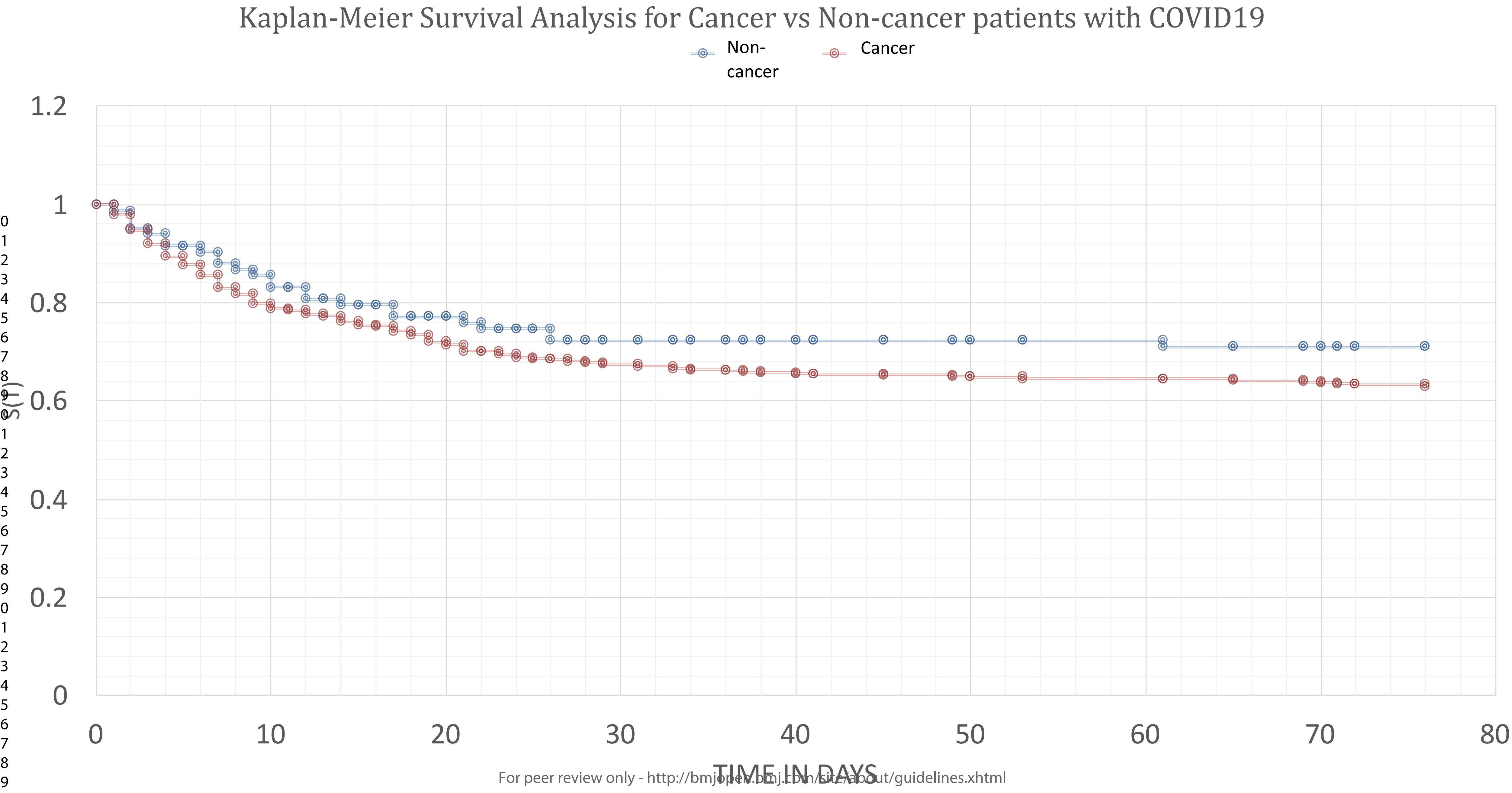
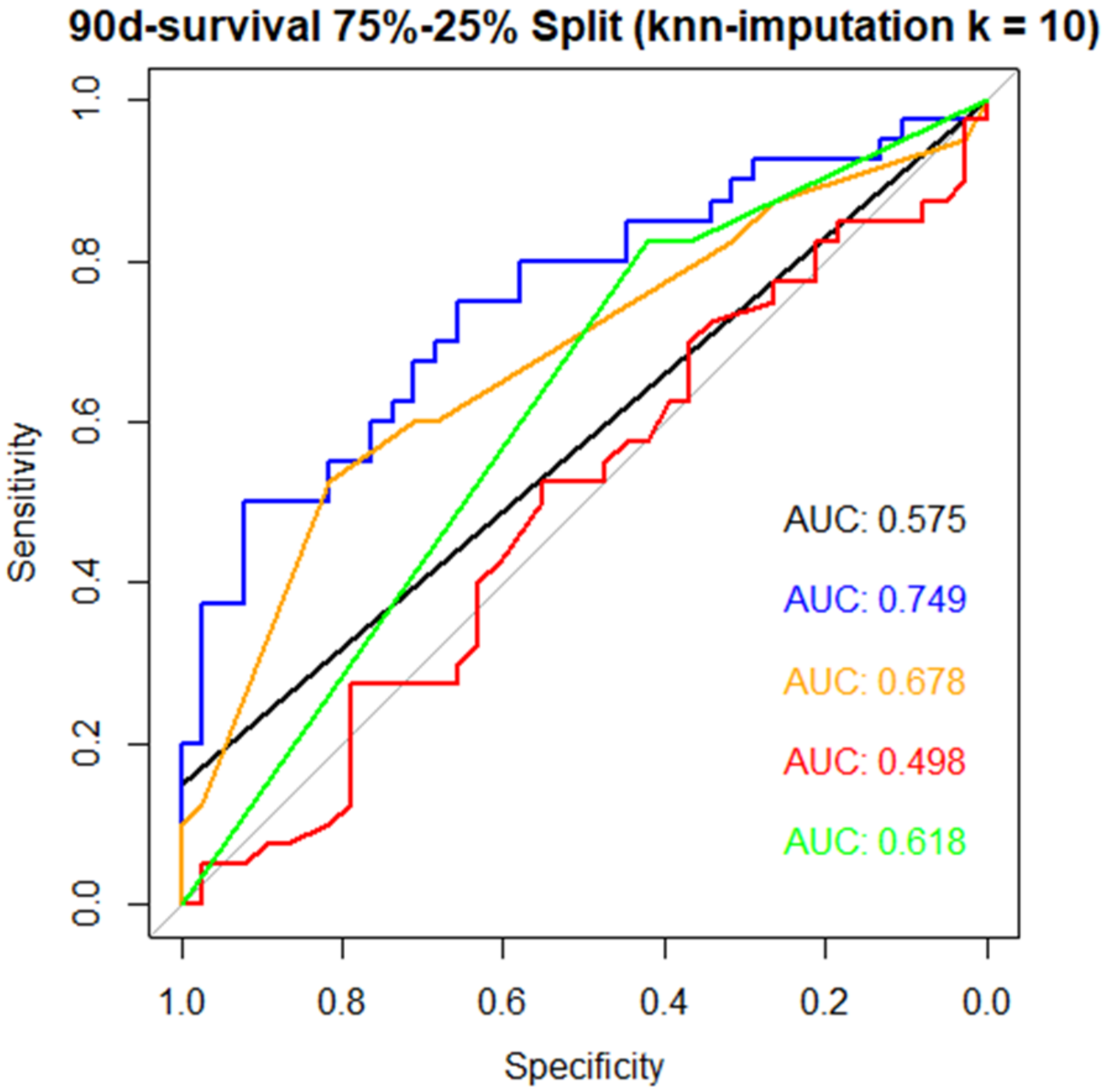


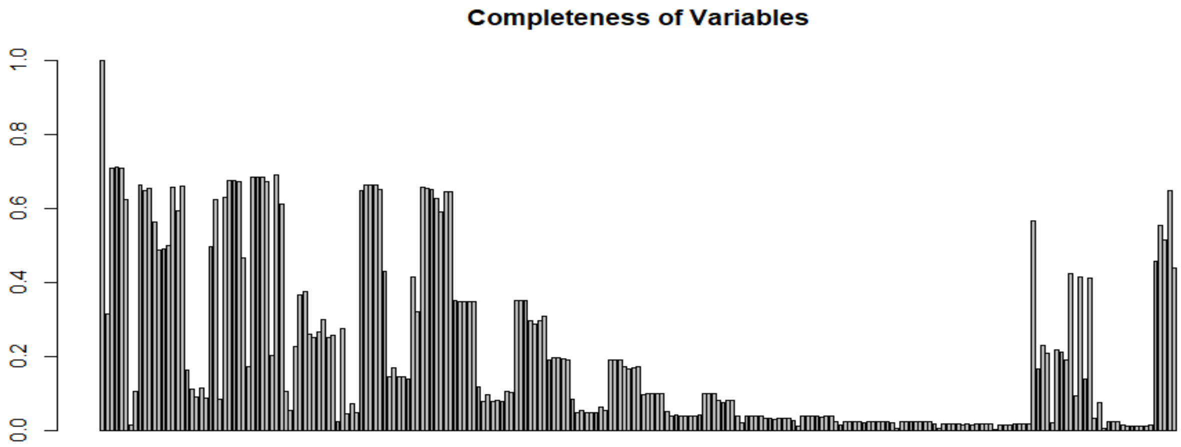
Figure 3: Relative importance of each variable in the machine learning algorithm in determining outcome from COVID-19 infection.


Kaplan-Meier Survival Analysis for Cancer vs Non-cancer patients with COVID19
Non-cancer
Cancer
S(T)
1.2
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
TIME IN DAYS


90d-survival 75%-25% Split (knn-imputation k = 10)
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC: 0.575
AUC: 0.749
AUC: 0.678
AUC: 0.498
AUC: 0.618

Supplementary Figure 1

The graph below plots the proportion of missing data for each variable:



If you only include variables with at least 60% completeness, there are 33 variables out of 233.

Below is a feature plot of each these variables against whether or not the patient was alive at 30 days (the only dependent variable with >= 60% completeness):

Supplementary Table 1

Univariate analysis of variables correlating with mortality risk in patients presenting with COVID-19. Age, haemoglobin (Hb), C reactive protein (CRP), urea, creatinine, glomerular filtration rate (GFR), low initial blood pressure (BP) were associated with the highest risk.

	Univariable			95% CI
	p	OR	Lower	Upper
Age	0.000	1.048	1.031	1.064
WCC...57	0.407	1.005	0.994	1.016
Hb...58	0.000	0.982	0.972	0.992
Plt...59	0.640	1.000	0.998	1.001
CRP...60	0.000	1.005	1.003	1.008
Ur...69	0.000	1.132	1.081	1.185
Cr...70	0.001	1.006	1.003	1.010
eGFR...71	0.000	0.977	0.967	0.988
ALT...72	0.377	1.002	0.997	1.007
Bili...74	0.019	1.034	1.005	1.062
Alb...75	0.533	0.993	0.973	1.014
Initial vital signs - Heartrate	0.567	0.997	0.989	1.006
Initial vital signs – BP, Mean, Arterial Pressure	0.048	0.987	0.973	1.000
Initial vital signs - RR	0.143	1.015	0.995	1.036
Initial vital signs - SpO_2	0.454	0.996	0.986	1.006
Sex	0.212	0.759	0.491	1.171
Ethnicity	0.994			
Smoking	0.647			
Cardiovascular disease	0.024	1.671	1.069	2.614
Diabetes Mellitus	0.253	1.321	0.819	2.130
Transplant patient	0.166	4.740	0.524	42.877
Cancer patient	0.024	1.832	1.082	3.101
Reason for admission	0.143	0.610	0.315	1.181
Main SX on admission - Cough	0.006	0.543	0.351	0.840
Main SX on admission – fever of 38C	0.039	0.631	0.408	0.976
Main SX on admission - Shortness of Breath	0.883	1.037	0.642	1.673

STROBE 2007 (v4) Statement—Checklist of items that should be included in reports of *cohort studies*

Section/Topic	Item #	Recommendation	Reported on page #
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	1
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2-3
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4
Objectives	3	State specific objectives, including any prespecified hypotheses	4
Methods			
Study design	4	Present key elements of study design early in the paper	5
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	5-8
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	5-8
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed	5-8
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	5-8
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	*5-8*
Bias	9	Describe any efforts to address potential sources of bias	5-8
Study size	10	Explain how the study size was arrived at	Not done
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	5-8
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	7-8
		(*b*) Describe any methods used to examine subgroups and interactions	7-8
		(*c*) Explain how missing data were addressed	7-8
		(*d*) If applicable, explain how loss to follow-up was addressed	7-8
		(*e*) Describe any sensitivity analyses	7-8
Results			

Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	8
		(b) Give reasons for non-participation at each stage	8-9
		(c) Consider use of a flow diagram	
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	8-9
		(b) Indicate number of participants with missing data for each variable of interest	8-9
		(c) Summarise follow-up time (eg, average and total amount)	8-9
Outcome data	15*	Report numbers of outcome events or summary measures over time	8-9
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	8-9
		(*b*) Report category boundaries when continuous variables were categorized	8-9
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	8-9
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	8-9
Discussion			10-13
Key results	18	Summarise key results with reference to study objectives	10-13
Limitations			
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	10-13
Generalisability	21	Discuss the generalisability (external validity) of the study results	10-13
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	3

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.